THE SECURITY FLAWS IN SOME AUTHENTICATION WATERMARKING SCHEMES

Yongdong Wu, Feng Bao and Changsheng Xu

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore, 119613
{wydong, baofeng, xucs}@i2r.a-star.edu.sg

ABSTRACT

Watermarking technology was originally proposed for copyright protection. Recently it has been applied to media authentication so that a proof of authenticity is inserted into the media instead of being appended to the media as a separated attachment. However, security requirements of the authentication are overlooked in some authentication watermark schemes. In this paper we analyze three authentication watermarking schemes and point out their security flaws. The first scheme is the color authentication scheme in [1]. The scheme is not secure in the sense that as long as an attacker obtains one authenticated image, he is able to forge authentic images without the secret key. The second scheme [2] is an authentication scheme but it is extended for ownership incorrectly. The third one, the robust invertible watermarking scheme [3], employs a multiple of secret random sequences to produce a watermark. However these sequences are independent of the original images, i.e., they remain invariable for different images. An adversary, having sufficient number of original images, can reconstruct the secret sequences by solving simultaneous equations. With these reconstructed sequences, the attacker can forge authentic image freely. The attack can be thwarted with content related sequences generated from both the secret key and the original image.

1. INTRODUCTION

From the viewpoint of a software engineer, digital content is no more than a 0’s and 1’s bit string. Any one with a handy editing tool can manipulate the digital content freely and easily. The modified content can be disseminated rapidly without any limitation. To protect copyright and/or content authentication/integrity, cryptography primitives such as hash functions and encryption/decryption algorithms are conventional tools. But the authentication message needs to be included in an extra attachment. Alternatively, digital watermarking approaches embed information into multimedia data to allow copyright enforcement or content authentication/integrity protection.

There are 3 classes of watermarking methods. A fragile watermarking scheme detects any manipulation to guarantee the content integrity. A semi-fragile scheme detects unacceptable image manipulations, while allowing lossy compression or other acceptable modifications. A robust watermarking scheme resists content modification to a highest degree. This method prevents an attacker from removing the watermark unless the quality of the content is reduced greatly.

In principle, any sort of image watermarking schemes can be used for content authentication that immunizes from forgery attack. The proposal of Byun et al. [1] replaces the Least Significant Bits (LSB) with the compression information of red and green colors as well as the logo of the owner. The verifier reconstructs the logo from the watermarked image so as to authenticate the original image. Wong et al. [2] provided 2 schemes for image authentication and ownership protection with the similar technology. These schemes not only detect the modification on the image but also locate manipulated blocks.

The fragile schemes [1,2] modify the LSB planes of the original image in an irreversible way. This is not satisfactory in some application fields such as medical images for legal reasons. An invertible authentication watermarking system tries to remove the distortion to recover the original image if the image is authentic. It provides new information assurance tools for integrity protection of sensitive imagery. Fridrich et al. [3] proposed an invertible watermarking method for authenticating digital images based on the robust watermarking scheme addressed in [4].

Unfortunately, the three above authentication paradigms cannot achieve the claimed security. Byun’s scheme [1] pays attention to the LSB only without caring about Most Significant Bits (MSB). Wong et al. overlook the forgery logo when their schemes are extended to ownership verification. Fridrich’s scheme is not secure either if the attacker can get sufficient number of pairs of
original images and watermarked images. In the following sections, we will elaborate their vulnerabilities and provide countermeasures for them.

The rest of the paper is organized as follows. We present the security analysis on the schemes [1,2,3] and the description of countermeasures to the attacks on them in sections 2, 3 and 4, respectively. At last, a concluding remark is provided.

2. COLOR IMAGE AUTHENTICATION SCHEME

2.1. Overview of Byun’s scheme

In ICME2002, Byun et. al [1] presented a fragile watermarking scheme for color image authentication aiming for that if even a single bit has been changed, the watermarked image is regarded as inauthentic. The embedding process is illustrated in figure 1. The color image includes 3 components: red $O_R$, green $O_G$ and blue $O_B$. In order to authenticate the image, LSB planes of $O_R$ and $O_G$ are compressed by a hash function $MD(\cdot)$. Hash value $H$ is XORed with a logo $W_0$ to produce $M$. $M$ is scrambled further with a private key2 of the verifier and ciphertext $E$ will be inserted into LSB of color $O_B$.

The watermark extracting process is the reverse of the above steps. The authors claimed that any change on a watermarked image will result in report of no watermark. Furthermore, the pixel changes of the watermarked image are reflected in the extracted watermark, which indicates the attacked area.

2.2. Attack and countermeasure

Although the scheme could extract the watermark and localize the attack positions correctly, the important image data—the MSBs of $O_R$, $O_G$ and $O_B$ are ignored in the embedding process. In other words, It just authenticates the LSBs. An attacker can change the MSBs of an authentic image, or replace LSBs of a non-authenticated image with LSBs of an authentic image to produce a bogus image. Because the LSB of the forged image is the same as the authentic one, a verifier will authenticate the forged image. Here, key1 for permutation can not provide any help. To defend against this counterfeiting attack, the input message of hash function $MD(\cdot)$ should comprise of MSB planes of $O_R$, $O_G$ and $O_B$ as well as the LSB planes of $O_R$ and $O_G$.

3. WONG’S AUTHENTICATION SCHEME

Wong et. al [2] described two authentication watermarking schemes: a secret key scheme and a public key scheme. The schemes are applicable to color image and greyscale images. For simplicity, we consider the security of the public scheme on greyscale images only. It is easy to extend the principle to secret scheme and color images. We partition original image $X$ into blocks. Let the $r^{th}$ block to be $X_r$ and $B_r$ to be a binary image used as the invisible watermark which is of the same size as $X_r$. If not, form by tiling periodically replicating to the desired size or append all zeros (or all ones).

3.1. Public watermarking scheme

![Fig. 2 Public watermarking insertion][2]

- **Input image** $X_r$
- **Block index** $r$
- **Image width** $M_r$
- **Image height** $N_r$
- **Image index** $I_r$
- **Watermark** $W_r$
- **Encryption key** $k$
- **Watermark block** $B_r$
- **Public Key** $E(k)$
- **Watermarked image** $X'_r$

The watermark extracting process is the reverse of the above steps. The authors claimed that any change on a watermarked image will result in report of no watermark. Furthermore, the pixel changes of the watermarked image are reflected in the extracted watermark, which indicates the attacked area.
Fig. 2 describes the watermark insertion procedure for each image block in Wong’s scheme. Firstly, the LSB bit plane of block $X_r$ is set to 0 to form $\tilde{X}_r$ then

$$W_r = B_r + H(I_r, M_s, N_s, r, \tilde{X}_r) \quad \cdots \quad (1)$$

$$S_r = E_k(W_r) \quad \cdots \quad (2)$$

$S_r$ is used to replace the LSB bit plane of $\tilde{X}_r$ to generate watermarked image $X_r^w$. At the retrieval side, $S_r$ and $\tilde{X}_r$ can be recovered from the watermarked image $X_r^w$ easily. Based on formulas (1) and (2), the watermark bitmap $B_r = W_r - H(I_r, M_s, N_s, r, \tilde{X}_r)$ which indicates the ownership of image $X$.

3.2. Analysis

In the authentication or ownership scheme, anyone can claim he is the owner of an image if he can prove his logo (e.g. trademark) to be the watermark. It is true that Wong’s scheme is viable for authentication purpose. However, the authors extended the scheme to ownership verification inappropriately. Specifically, they inserted a visible watermark and then inserted an invisible authentication watermark to the visibly watermarked image. Because the watermarking scheme is fragile, the pirate can replace the visible watermark with his own logo, and execute the above watermarking process with his own private key and bitmap logo. This piracy indicates a new owner but the quality is almost the same as the original one.

Similarly, the secret scheme can not be used for ownership verification either.

4. INVERTIBLE WATERMARKING SCHEME

Watermarking schemes have to change the original image so as to embed the watermark. Generally, this distortion is not reversible even though it is imperceptible. A novel watermarking scheme named lossless watermarking method was proposed to avoid the distortion in the recent years. In the following sections, we analyze one of the lossless watermarking authentication scheme [3].

4.1. Invertible authentication scheme

In this scheme, the embedder shares a secret key with the verifier. The secret key is used as a seed to produce a multiple of random sequences. Afterwards, some subsequences of them are selected based on the hash value of original image $X$ to form a watermark. This watermark is embedded into $N$ DCT coefficients of the original image to obtain watermarked image $X'$. The following steps introduce the embedding process in Fridrich scheme [3].

- Generate hash value $h = (h_0, h_1, \ldots, h_{24})$ of original image $X$ with a cryptographic hash function (e.g MD5). Where the size of $h_i$ is 6 bits.
- Generate 25 pseudo-random sequences $\xi(0), \xi(1), \ldots, \xi(24)$ uniformly distributed in $[-1, 1]$ with the secret key. Segment each pseudo-random sequence $\xi(i)$ into 64 sub-sequences whose length is $N$. Therefore there are 1600 sub-sequences $\eta(i,s) = \xi(s_0), \ldots, \xi(s_{62})$, where $0 \leq i \leq 24$ and $0 \leq s \leq 63$.
- Represent hash value $h$ as a summation

$$W = \left( \frac{3}{25} \sum_{s=0}^{24} \eta(i, h_i) \right) \quad (3)$$

- Calculate the DCT coefficients of the whole image and modulate DCT coefficients $D_k(k=0,1,\ldots,N-1)$ to $D_k = D_k + \alpha W_k$. Where $\alpha$ is a constant.
- Watermarked image $X'$ is obtained by performing the inverse DCT using the modulated coefficients $D_k$.

To verify an image, a verifier who knows the secret reconstructs sequences $\xi(i)$ ($0 \leq i \leq 24$). Then, he calculates the correlation value between coefficients $D_k$ and each sub-sequences $\eta(i,s)$. Selects the sub-sequence whose correlation value is the maximum among each sequence $\xi(i)$. These sub-sequences, representing the hash value of the original image, can be used to recover the original image if there is no tampering. Therefore, if the reconstructed hash value is identical to the new hash value generated from the reconstructed image, the verifier can confirm the authenticity of watermarked image $X'$.

4.2. Cover-image attack

Fridrich [5] pointed out that cover-image attack requires multiple pairs of original images and watermarked images in advance. This assumption is reasonable to some extent. For example, an attacker can somehow get access to the raw images such as out-of-date images. The well-known Kerckhoff principle [6,8] indicates that a strong security system should publish all information with exception of the keys. That is to say, the system security should not depend on a confidential image database.

In the authentication scheme [3], the sequences are uniquely determined by the secret key, and invariable for all authenticated images. However, if an adversary has plenty of original images and corresponding watermarked images, he can set up a series of simultaneous equations. By solving them, the sequences determined by the secret are reconstructed. These sequences enable the adversary to forge authentic image at will. The following steps illustrate how to find the sequences given sufficient
number of original images and corresponding watermarked images.
• Select \(\tau\)th pair of original image \(X\) and watermarked image \(X'\) randomly.
• Calculate DCT coefficients \(D_i\) and \(D_i'\) of the original image and the watermarked image respectively, then
  \[W_k = (D_k - D_k')/\alpha, \ k = 0, 1, \ldots, N-1.\]
  From formula
  \[
  W_k = \left( \sum_{i=0}^{24} \eta(i,s_i) \right) - \frac{25}{3\alpha^2} (D_0 - D_0', D_1 - D_1', \ldots, D_{N-1} - D_{N-1})
  \]

  \[
  \eta(i,s_i) = \frac{25}{3\alpha^2} (D_0 - D_0', D_1 - D_1', \ldots, D_{N-1} - D_{N-1})
  \]

  \[\cdots \quad (3)\]

  
  • Compute hash value \(h = (s_0, s_1, \ldots, s_{24})\) of original image \(X\), where the size of \(s_i\) is 6 bits. For clarity, let
  \[x_i = \eta(i,s_i), 0 \leq j = 64i + s_i \leq n = 1599,\]
  and
  \[T = \frac{25}{3\alpha^2} (D_0 - D_0', D_1 - D_1', \ldots, D_{N-1} - D_{N-1}).\]

  Rewriting formula (4) into
  \[b_{00} b_{01} \cdots b_{0(t-1)} b_{0n} \]
  \[b_{10} b_{11} \cdots b_{1(n-1)} b_{1n} \]
  \[\vdots \quad \vdots \quad \vdots \]
  \[b_{(t-1)0} b_{(t-1)1} \cdots b_{(t-1)(n-1)} b_{(t-1)n} \]
  \[b_{r0} b_{r1} \cdots b_{r(t-1)} b_{rn} \]
  \[
  \begin{pmatrix}
  x_0 \\
  x_1 \\
  \vdots \\
  x_{n-1} \\
  x_n
  \end{pmatrix}
  =
  \begin{pmatrix}
  T_0 \\
  T_1 \\
  \vdots \\
  T_{n-1} \\
  T_n
  \end{pmatrix}
  \]

  \[\cdots \quad (5)\]

  If the rank of coefficient matrix of formula (5) is \(n, x_j\)
  \((j=0, 1, \ldots, n-1)\) can be reconstructed. That is to say, the sub-sequences are recovered. After obtain the sequences, the attack can generate the forgery images at will. The forgery procedure includes the last 3 steps of the embedding method of Fridrich’s scheme [3].

4.3. Countermeasure

The above attack exploits the weakness that sequences are fixed in embedding any images. If we can produce content related sequences, the weakness will be removed. Considering that some DCT coefficients are not modified, we can generate a PRNG seed with the secret and the unmodified coefficients to produce content-based sequences. Because the new sequences used for watermarking are image-dependent, the simultaneous equations (5) are hard to set up, therefore the above attack can not be mounted.

Alternatively, the sequences can be generated with the secret and other recoverable information, for example, the relationship between the DCT coefficients mentioned in semi-fragile watermarking scheme [7].

5. CONCLUSION

In this paper, we analyze 3 image authentication watermarking schemes. The first one has one shortcoming that the authentication is valid only for part of the image data. Of course, this flaw can be avoided easily when all content to be protected is involved in the authentication scheme. The second one has no obvious authentication weakness, but the authors misused it for ownership verification. The third one is an invertible authentication scheme. It is vulnerable to cover-image attack by reconstructing sequences. Although the attacker does not know the secret, he can forge authentication images with the reconstructed sequences. To foil this cover-image attack, we suggest creating content-based sequences which depend on the original image.

6. REFERENCES